УДК 621.391

Е.Г. ЖИЛЯКОВ, д-р техн. наук, проф., зав. каф., НИУ "БелГУ", Белгород,
А.А. ФИРСОВА, ассистент, НИУ "БелГУ", Белгород

СЕГМЕНТАЦИЯ РЕЧЕВЫХ СИГНАЛОВ НА ОСНОВЕ СУБПОЛОСНОГО АНАЛИЗА

Введены понятия нормированной субполосной корреляции и субполосного расстояния. Предложен новый метод сегментации речевых сигналов по границам звуков речи, основанный на использовании субполосного расстояния. Предложен новый метод выделения отрезков речевых сигналов, порождаемых звуками речи с почти периодической структурой. Библиогр.: 8 назв.

Ключевые слова: нормированная субполосная корреляция, субполосное расстояние, сегментация речевых сигналов, звуки речи с почти периодической структурой.

Постановка проблемы. В настоящее время возрастает интерес к разработке разнообразных речевых технологий [1, 2], в том числе к созданию методов и алгоритмов автоматического распознавания речи [3]. Обработке при этом подвергаются речевые сигналы (РС), которые являются результатами регистрации значений электромагнитных колебаний на выходе микрофонов при воздействии акустических колебаний на их входах, возникающих в результате речевого обмена. Одной из важных задач является сегментация РС, то есть разбиение их на отрезки, которые порождаются разными звуками речи или паузами речи. В данной работе эта задача рассматривается без идентификации порождающих акустические колебания звуков речи.

Основные рассматриваемые аспекты: обнаружение переходов речь/пауза и пауза/речь; моменты смены одного звука речи другим; выделение отрезков РС, порождаемых звуками речи с почти периодичной структурой, прежде всего вокализованных.

Анализ литературы. Эффективность алгоритма сегментации определяется точностью определения границы между различными звуками. Существующие методы сегментации речевых сигналов по звукам речи можно разделить на несколько классов: основанные на использовании спектрального анализа в базисе Фурье; основанные на использовании вейвлет-анализа; основанные на использовании коэффициентов корреляции. Отдельным классом можно выделить методы, основанные на различиях энергетических характеристиках, оценка которых осуществляется во временной области [4 – 6].

© Е.Г. Жиляков, А.А. Фирсова, 2013

Существующие методы сегментации либо позволяют определять только границы слогов или предложений, либо приводят к появлению дополнительных границ на участках, соответствующих одному звуку.

В основе многих из разработанных подходов используются частотные представления, так как порождаемые звуками речи отрезки РС обладают свойством концентрации энергии в достаточно "узких" полосах частотной оси. В связи с этим можно упомянуть рассматриваемое в литературных источниках разбиение частотной полосы на так называемые критические полосы слуха, которые опосредованно отражаются на частотных свойствах РС. Таким образом, адекватным подходом к обработке РС является субполосный анализ, когда их свойства соотносятся с некоторым разбиением оси частот на интервалы конечной ширины. Причем, в виду зависимости частотного распределения энергий от вида произносимого звука, анализу необходимо подвергать отрезки РС конечной длительности.

Цель статьи – разработка методов сегментации РС по границам звуков и выделение отрезков РС, порождаемых звуками речи с почти периодичной структурой, на основе субполосного анализа.

Основы субполосного анализа РС. Предполагается, что РС представлены эквидистантными отсчетами

$$x_i = x(i\Delta t),\ \ i = 1, 2, \ldots,$$

с частотой дискретизации:

$$f_д = 1/\Delta t \geq 8000\ Гц.$$

Известно [1, 3], что все звуки русской речи обладают свойствами концентрации энергии в пределах малой доли частоты дискретизации. Поэтому адекватным подходом к сегментации РС является применение субполосного анализа, когда их характеристики соотносятся некоторым разбиением области нормированных частот [7]

$$-\pi \leq \omega \leq \pi, \tag{1}$$

на частотные интервалы:

$$\Omega_r = [-\Omega_{2r}, -\Omega_{1r}] \vee [\Omega_{1r}, \Omega_{2r}], \tag{2}$$

где $r = 1, \ldots, R$;

$$\Omega_{2r} > \Omega_{1r};\ \ \Omega_{2r} \leq \pi. \tag{3}$$

В соответствии с конечной длительностью звуков речи анализу должны подвергаться конечные наборы отсчетов РС (векторы)

$$\vec{x}_N = (x_1, ..., x_N)^T , \tag{4}$$

где T – символ транспонирования.

Положим

$$X_N(\omega) = \sum_{i=1}^{M} x_i e^{-j\omega(i-1)} . \tag{5}$$

Имеет место формула обращения [8]

$$x_k = \int_{-\pi}^{\pi} X_N(\omega) e^{j\omega(k-1)} d\omega / 2\pi , \tag{6}$$

и справедлива формула Парсеваля

$$g_{xy} = (\vec{x}_N \vec{y}_N) = \sum_{i=1}^{M} x_i y_i = \int_{-\pi}^{\pi} X_N(\omega) Y_N^*(\omega) d\omega / 2\pi , \tag{7}$$

где звездочка означает комплексное сопряжение, а $\vec{y}_N = (y_1, ..., y_N)^T$.

В частности, имея в виду частотные интервалы (2), соотношение (7) можно переписать в виде суммы:

$$(\vec{x}_N, \vec{y}_N) = \sum_{i=1}^{R} G_r(\vec{x}_N, \vec{y}_N) , \tag{8}$$

слагаемые которой

$$G_r(\vec{x}_N, \vec{y}_N) = \int_{\omega \in \Omega r} X_N(\omega) Y_N^*(\omega) d\omega / 2\pi \tag{9}$$

естественно называть субполосными корреляциями.

Кроме того, можно ввести понятие частей энергии, попадающих в частотные интервалы

$$\left\|\vec{x}_N\right\|^2 = \sum_{r=1}^{R} P_r(\vec{x}_N); \quad \left\|\vec{y}_N\right\|^2 = \sum_{r=1}^{R} P_r(\vec{y}_N), \tag{10}$$

где

$$P_r(\vec{z}_N) = \int_{\omega \in \Omega_r} \left|Z_N(\omega)\right|^2 d\omega / 2\pi , \tag{11}$$

где $\vec{Z}_N(\omega)$ – трансформанта Фурье вектора $\vec{z}_N$, и понятие субполосных нормированных корреляций

$$\rho_r(\vec{x}_N, \vec{y}_N) = \frac{G_r(\vec{x}_N, \vec{y}_N)}{\sqrt{P_r(\vec{x}_N) P_r(\vec{y}_N)}} . \tag{12}$$

Они, очевидно, удовлетворяют неравенству

$$\left|\rho_r(\vec{x}_N, \vec{y}_N)\right| \le 1. \quad (13)$$

Именно характеристики (11) и (12) в дальнейшем положены в основу разрабатываемых алгоритмов сегментации речевых сигналов.

Примечательно, что для их вычисления нет необходимости переходить в частотную область, т.к. подстановка в представление (9) и (11) определений вида (5) позволяет получить реализуемые непосредственно во временной области билинейные и квадратичные формы

$$G_r(\vec{x}_N, \vec{y}_N) = \vec{x}_N^T A_r \vec{y}_N, \quad (14)$$

$$P_r(\vec{z}_N) = \vec{z}_N^T A_r \vec{z}, \quad (15)$$

где A_r субполосная матрица с элементами:

$$a_{ik}^r = \{\sin(\Omega_{2r}(i-k)) - \sin(\Omega_{1r}(i-k))\} / \pi(i-k). \quad (16)$$

Отметим, что соотношения (14) и (15) позволяют вычислить точные значения частей энергии отрезков сигналов, приходящихся на заданный частотный интервал, и соответствующих субполосных корреляций.

Селекция пауз в речевых воздействиях

Исходная (нулевая) гипотеза формируется следующим образом.

H_0: отрезок сигнала $\vec{x}_N$ зарегистрирован в паузе речи так что

$$\vec{x}_N = \vec{u}_N, \quad (17)$$

где $\vec{u}_N = (u_1, ..., u_N)^T$ – вектор отрезков шумов.

Альтернатива имеет следующую формулировку.

H_1: хотя бы часть отчетов зарегистрирована в присутствии речевого воздействия, которые аддитивно взаимодействуют с шумом, то есть

$$\vec{x}_N = \vec{z}_N + \vec{u}_N, \quad (18)$$

где $\vec{z}_N = (z_1, ..., z_N)^T$ - вектор отчетов возбуждаемых речью, часть из которых может быть равна нулю.

В качестве решающей функции предлагается использовать:

$$F_r(\vec{x}) = \frac{P_r(\vec{x}_N)}{E[P_r(\vec{u})]}, \quad (19)$$

где E – символ математического ожидания.

Гипотеза H_0 отвергается при выполнении следующего неравенства

$$\max F_r(\vec{x}) > h_\alpha, \quad (20)$$

где максимум определяется для всех частотных интервалов, а h_α – некоторый порог.

Предполагается, что имеется возможность предварительного обучения, на этапе которого при заведомом отсутствии речи можно определить оценки $\overline{P}_r(\vec{u}_N)$ математических ожиданий частей энергий шумов и оценку величины порога в (20), удовлетворяющую условию:

$$PR\left\{ P_r(\vec{u}_N) \ / \overline{P}_r(\vec{u}_N)) \geq h_\alpha \right\} \leq \vec{\alpha}, \qquad (21)$$

где PR – оценка вероятности, α – желаемый уровень вероятности ошибок первого рода, а $\vec{\alpha}$ его оценка при использовании оценок математических ожиданий.

Оценивание математических ожиданий и порога можно осуществить по одному и тому же достаточно большому количеству отчетов шумов, при отсутствии речи, например по 10 000 отрезкам необходимой длительности (порядка 1,5 секунд).

Предполагая для простоты, что шумы в паузах, являются гауссовыми с независимыми отсчетами, причем

$$E[u_t] = 0, \qquad (22)$$

$$\sigma_0^2 = E[u_t^2], \qquad (23)$$

можно показать справедливость следующих соотношений

$$m_r = E[P_r(\vec{u})] = \sigma_u^2 N(\Delta\Omega_r / \pi)^2, \qquad (24)$$

$$\sigma_r^2 = E[(P_r(\vec{u}) - m_r)^2] = 2N\sigma_u^4(\Delta\Omega_r / \pi)^2, \qquad (25)$$

где $\Delta\Omega_r = \Omega_{2r} - \Omega_{1r}$.

Таким образом, имеет место

$$\gamma_r = \sigma_r / m_r = 2/(N)^{1/2}. \qquad (26)$$

То есть в отсутствии сигнала дисперсия решающей функции (19) обратно пропорциональна длительности обрабатываемого отрезка, а ее математическое ожидание равно единице.

В виду нестационарности речевых воздействий исследовать мощность критерия (20) (левая часть) не представляется возможным. Отметим только, что использование максимального значения решающей функции, по крайней мере, в случае белого шума, позволяет в среднем эффективно отреагировать на появления дополнительной энергии, которая сосредоточена в малой доле частотной полосы.

Сегментация квазипериодических звуков речи

Некоторые звуки речи порождают отрезки речевых сигналов с достаточно отчетливой повторяемостью фрагментов, которые естественно называть квазипериодами, наиболее отчетливо это свойство выражено при произнесении так называемых вокализованных звуков, к которым относятся гласные.

Выделение участков квазипериодичности является важной задачей обработки речевых сигналов, о чем свидетельствуют многие работы, в которых она рассматривается (см. например [1, 3]). При этом основное внимание уделяется оцениванию периода, так называемого, основного тона. Рассмотрим некоторые проявления свойств периодичности.

Пусть отрезок отсчетов сигнала $\vec{x} = (x_1, \ldots, x_N)^T$ обладает свойством периодичности

$$x_{i+kM} = x_i \tag{27}$$

и для простоты имеет место

$$N = L \times M\ , \tag{28}$$

где L и k – целые числа.

Тогда для его трансформанты Фурье справедливы соотношения

$$|X(\omega)|^2 = D^2(\omega) \times \left|X_M(\omega)\right|^2, \tag{29}$$

$$\left|X_M(\omega)\right|^2 = \left|\sum_{k=1}^{M} x_k e^{-j\omega(k-1)}\right|^2, \tag{30}$$

$$D^2(\omega) = \sin^2\left(LM\frac{\omega}{2}\right) / \sin^2\left((M\frac{\omega}{2}\right). \tag{31}$$

Функция (31) в частотной области имеет максимумы в точках (кроме $\omega = 0$)

$$\omega_k = k\frac{2\pi}{M};\ k = 1, \ldots, n\ , \tag{32}$$

так что первый максимум соответствует периоду.

Если, в свою очередь, и произведение (29) будет иметь в точке $2\pi / M$ наибольший из экстремумов, то это позволяет оценить величину периода. Однако сомножитель (30) может иметь максимумы в других точках оси частот, так что наибольшее значение (29) достигается также в другой точке, чем определяется (32) при $k = 1$. Такой эффект, например, проявляется в том, что достаточно эффективный метод оценивания периода основного тона на основе вычисления скалярных произведений

$$G(\tau) = \sum_{i=1}^{N} x_i x_{i+\tau}, \tag{33}$$

и использования оценки

$$\overline{M} = \arg\max_{1\le\tau\le 0} G_r(\tau) \tag{34}$$

также дает заниженные значения длительности периода основного тона.

Вместе с тем наличие квазипериодичности будет в той или иной мере проявляться во всех точках вида (32), так как сомножитель (30) будет отличен от нуля на всей частотной оси. Поэтому представляется естественным перейти к субполосным коэффициентам автокорреляции

$$\rho_r(\tau) = G_r(\tau) / \sqrt{P_r(\vec{x}_N) P_r(\vec{x}_N^{\tau})}, \tag{35}$$

где $\vec{x}_N^{\tau} = (x_{1+\tau}, .., x_{N+\tau})^T$;

$$G_r(\tau) = \vec{x}_N^T A_r \vec{x}_N^t. \tag{36}$$

При этом в качестве оценки периода предлагается использовать:

$$\overline{M} = \arg\max_{1<\tau\le L} \sum_{r=1}^{R} \ln \frac{1+\rho_r(\tau)}{1-\rho_r(\tau)} / R. \tag{37}$$

Используемое здесь усреднение преобразования Фишера повышает устойчивость оценки.

Определение границ между звуками речи

Основная гипотеза имеет следующий вид:

H_0: отрезки речевого сигнала $\vec{x} = (x_1, \ldots, x_N)^T$ и $\vec{x}_N^N = (x_{N+1}, \ldots, x_{2N})^T$ порождены одним и тем же звуком речи.

Положим

$$d_n(\vec{z}_N) = P_r(\vec{z}_N) / \left\|\vec{z}_N\right\|^2, \tag{38}$$

и введем на основе этих долей энергий понятие субполосного расстояния

$$\begin{aligned} V_N &= \Big(\sum_{r=1}^{R} \left((d_r(\vec{x}_N))^{1/2} - (d_r(\vec{x}_N^N))^{1/2}\right)^2 / 2\Big)^{1/2} = \\ &= \Big(1 - \sum_{r=1}^{R} (d_r(\vec{x}_N) d_r(\vec{x}_N^N))^{1/2}\Big)^{1/2}. \end{aligned} \tag{39}$$

Для проверки исходной гипотезы предлагается использовать решающую функцию

$$W_N = s_1^2 / s_2^2 \times V_N , \qquad (40)$$

где

$$s_1^2 = \max\left\{\left\|\vec{x}_N\right\|^2, \left\|\vec{x}_N^N\right\|^2\right\};$$

$$s_2^2 = \min\left\{\left\|\vec{x}_N\right\|^2, \left\|\vec{x}_N^N\right\|^2\right\}.$$

Гипотеза отвергается при выполнении неравенства

$$W_N > \vartheta_\alpha , \qquad (41)$$

где ϑ_α – порог, который соответствует некоторой желаемой вероятности ошибок первого рода.

Выводы. В результате проделанной работы был предложен новый метод селекции отрезков РС, порождаемых квазипериодическими звуками русской речи, основанный на введенном в работе понятии нормированной субполосной корреляции. Данный метод позволяет осуществлять селекцию отрезков РС, порождаемых квазипериодическими звуками русской речи с вероятностью ошибочного принятия решения не более 0,12, а также определить значение периода основного тона с погрешностью не более 0,02. Разработана новая решающая функция сегментации РС по границам звуков речи, основанная на использовании введенного в работе понятия субполосного расстояния. Предложенная решающая функция позволяет определять границы звуков русской речи с вероятностью ошибочного принятия решения не более 0,15÷0,20.

Работа выполнена в рамках проекта № 8.2251.2011 Государственного задания Министерства образования и науки РФ подведомственным ВУЗам на выполнение НИР в 2013 году.

Список литературы: **1.** *Шелухин О.И.* Цифровая обработка и передача речи / *О.И. Шелухин, Н.Ф. Лукьянцев.* – М.: Радио и связь, 2000 – 456 с. **2.** *Арлазаров В.Л.* Речевой ввод/вывод как развитие человеко-машинных интерфейсов / *В.Л. Арлазаров* // Информационные технологии и вычислительные системы. – 2004. – № 2. – С. 3-10. **3.** *Сорокин В.Н.* Структура проблемы автоматического распознавания речи / *В.Н. Сорокин* // Информационные технологии и вычислительные системы. – 2004. – № 2. – С. 25-40. **4.** *Сорокин В.Н.* Сегментация речи на кардинальные элементы / *В.Н. Сорокин, А.И. Цыплихин* // Информационные процессы. – 2006. – Т. 6. – № 3. – С. 177-207. **5.** *Дремин И.М.* Вейвлеты и их использование / *И.М. Дремин, О.В. Иванов, В.А. Нечитайло* // Успехи физических наук. – 2001. – Т. 171. – № 5. – С. 465-500. **6.** *Ермоленко Т.Н.* Алгоритмы сегментации с применением быстрого вейвлет-преобразования

/ *Т.Н. Ермоленко, В.И. Шевчук* // Статьи, принятые к публикации на сайте международной конференции Диалог'2003. www.dialog-21.ru. **7.** Вариационные методы анализа сигналов на основе частотных представлений / *Е.Г. Жиляков, С.П. Белов, А.А. Черноморец* // Вопросы радиоэлектроники, серия ЭВТ. – 2010. – Вып. 1. – С. 10-26. **8.** *Рабинер Л.* Теория и применение цифровой обработки сигналов / *Л. Рабинер, Голд Бю.* – М.: Мир, 1978. – 848 с.

Поступила в редакцию 03.04.2013

УДК 621.391

Сегментація мовних сигналів на основі субполосного аналізу / Жиляков Є.Г., Фірсова А.А. // Вісник НТУ "ХПІ". Серія: Інформатика та моделювання. – Харків: НТУ "ХПІ". – 2013. – № 39 (1012). – С. 73 – 81.

Введено поняття нормованої субполосної кореляції і субполосної відстані. Запропоновано новий метод сегментації мовних сигналів на кордонах звуків мови, що засновані на використанні субполосного відстані. Запропоновано новий метод виділення відрізків мовних сигналів, породжуваних звуками мови з майже періодичною структурою. Бібліогр.: 8 назв.

Ключові слова: нормована субполосна кореляція, субполосна відстань, сегментація мовних сигналів, звуки мови з майже періодичною структурою.

UDC 621.391

Segmentation of speech signals based on analysis sub-banding / Zhilyakov E.G., Firsov A.A. // Herald of the National Technical University "KhPI". Subject issue: Information Science and Modelling. – Kharkov: NTU "KhPI". – 2013. – №. 39 (1012). – P. 73 – 81.

The concepts of the normalized sub-banding correlation and sub-banding distance. A new method for segmentation of speech signals on the limits of the sounds of speech, based on the use sub-banding distance. A new method for isolation of segments of speech signals generated by the sounds of speech from the post of a periodic structure. Refs.: 8 titles.

Keywords: normalized sub-banding correlation sub-banding distance, segmentation of speech signals, the sounds of speech with a beinah periodic structure.